A Weakly Supervised Learning Technique for Classifying Facial Expressions

S L Happy, Antitza Dantcheva, Francois Bremond

A Weakly Supervised Learning Technique for Classifying Facial Expressions

S L Happy, Antitza Dantcheva and Francois Bremond
Inria Sophia Antipolis - Méditerranée, 06902, France

***Abstract*— The universal hypothesis suggests that the six basic emotions - anger, disgust, fear, happiness, sadness, and surprise - are being expressed by similar facial expressions by all humans. While existing datasets support the universal hypothesis and comprise of images and videos with discrete disjoint labels of profound emotions, real-life data contains jointly occurring emotions and expressions of different intensities. Models, which are trained using categorical one-hot vectors often over-fit and fail to recognize low or moderate expression intensities. Motivated by the above, as well as by the lack of sufficient annotated data, we here propose a weakly supervised learning technique for expression classification, which leverages the information of unannotated data. Crucial in our approach is that we first train a convolutional neural network (CNN) with label smoothing in a supervised manner and proceed to tune the CNN-weights with both labelled and unlabelled data simultaneously. Experiments on four datasets demonstrate large performance gains in cross-database performance, as well as show that the proposed method achieves to learn different expression intensities, even when trained with categorical samples.**

I. Introduction

Understanding human emotions is pertinent for the associated benefits in applications such as human-computer interaction, healthcare, surveillance and driver safety. Facial expression recognition (FER) aims at inferring emotions based on visual cues from face images. One major limitation in FER remains the lack of sufficient annotated data. Manual annotation of expressions is subjective and time-consuming, as well as notably impeded by different subjects (i.e., inter-individual variation) and different intensity-degrees within expressions (i.e., intra-individual variation).

Convolutional neural networks (CNNs) [1] have been efficiently utilized in many machine learning applications including object detection, image enhancement, speech analysis, natural language processing, representing the current state-of-the-art of such applications. Recently, FER - approaches based on CNNs [2], [3], [4], [5], [6] have attempted to replace classical approaches based on handcrafted features. As the majority of expression databases contain images or videos in the magnitude of few hundreds, this *limited dataset-size* poses a severe challenge in training of full-fledged CNNs. An additional FER - challenge concerns the related databases, which contain images or videos capturing subjects, exhibiting *discrete emotion categories* of *high intensity*. Models, trained on such data are predestined to fail when tested with real-life-data, where e.g., expressions of low-intensity occur frequently.

In the case of limited dataset-size, unannotated or weakly-annotated samples have been used in *weakly supervised methods* [7], [8], [9], achieving performances comparable to these of models, trained with a large labelled dataset. In weak supervision scenarios, a portion of training data might not be annotated or wrongly annotated [10]. As previously noted, in FER the annotation is categorical (i.e., irrespective of the expression intensity), as well as subjective. Given the above constraints and challenges, weakly supervised techniques offer to provide a good solution, taking advantage of unannotated data.

Contributions

Motivated by the above, we here propose a novel FER method which addresses both, the limited data-size, as well as lack of expression-intensity annotation in the same framework by incorporating both, transfer learning and weakly supervised learning. Firstly, we train an initial CNN model using the limited labelled data, employing the pre-trained weights of VGG-Face [11]. Label smoothing is applied to prevent the model from generating high confidence scores. Next, the model is updated using both labelled and unlabelled data. Specifically, a fraction of labelled data is bootstrapped with high confidence unlabelled data to update the model. Subsequently, the prediction scores of the current model are used as ground-truth distribution of unlabelled data for the next model update, by using both labelled and unlabelled data. In addition, we regulate the prediction confidence of the model on labelled data in order to have a prediction confidence higher than certain threshold for supervised data.

In summary, the contributions of the paper is two-fold. Firstly, we propose a weakly supervised technique to train a CNN model using both, labelled and unlabelled samples simultaneously, leveraging on the information available in large unannotated data. Secondly, we demonstrate that expression-intensity can be learned from the data annotated with discrete expression categories. The proposed method achieves a significant improvement in cross-database experiments.

II. Related Work

While methods based on hand-crafted features dominated the performance in FER [12] for a long time, recently CNNs have replaced such methods. Jung *et al.* [2] attempted to encode temporal appearance and geometry features in a CNN

This work was supported by the French Government (National Research Agency, ANR) under grant agreement ANR-17-CE39-0002.

framework. Boosted Deep Belief Network [3] improved the expression recognition performance by jointly learning the feature representation, feature selection, and classification. Peak-piloted deep network [4] implicitly embedded facial representation of both, peak and non-peak expression frames. Identity-aware CNN [6] jointly learned expression and identity related features to improve person independent recognition performance.

Addressing the above described *limited data-size* problem, Oquab *et al.* [13] proposed the transfer of mid-level image representation for related source and target domains. The authors showed that transfer of network parameters learned on large-scale annotated data can significantly improve the performance of a task with limited amount of training data. For example, improved performance was observed in emotion classification [14] using the transfer of network weights trained on ImageNet. Similarly FaceNet2ExpNet [5] fine-tuned the FaceNet in order to capture high level expression semantics.

Weakly supervised network training methods were reported in the literature as an additional solution for datasets of limited size. Unsupervised clustering (for example, K-means clustering [15]) provided an intuition that data distribution can be leveraged towards improving model performance. Another well received technique relates to the training of stacked auto-encoders with unlabelled data, further improved using supervised methods [16], [17], [18]. Moreover, the use of adversarial networks, learning from abundant unlabelled data [19], [20] has been very well accepted. Decoupled deep neural network [9] involve two separate networks for classification and segmentation of objects. Apart from these, self-training or incremental semi-supervised learning methods [21], [7], [8] have been well studied. In this context an initial model was built from the limited labelled data and it was used to predict the scores of the unlabelled data. Unlabelled data with high confidence scores was considered as the ground truth and utilized to further train the model. Similar semi-supervised or weakly supervised approaches were used in speech processing [22].

Rosenberg *et al.*[21] proposed a self-learning object detection model, where the *weakly label* refers to the data in which the probability of presence of an object in the image is provided instead of its exact location in that image. In case of image segmentation, weak annotation refers to the availability of object labels as bounding boxes, where the pixel-wise strong annotation is missing [9]. In this manuscript, we refer to the expression ground-truth provided in the database as weak annotation, as manual labeling of expressions is subjective. Using weak annotations, we encourage the CNN to learn the accurate expression representation with respect to their intensities.

III. Proposed Method

Given the input image x, the classification network learns to accurately predict the relevance expression scores $p(k|x,\theta)$, where θ are the network parameters and $k \in \{1,2,...,K\}$ represents K classes. For a soft-max layer, we have $p(k|x,\theta) = \frac{\exp(z_k)}{\sum_{i=1}^{K}\exp(z_i)}$, where z_i are the unnormalized log probabilities. In supervised learning scenarios, the ground-truth distribution $q(k|x)$ is used to train the network parameters (θ) by minimizing the cross-entropy loss function

$$l = -\sum_{k=1}^{K} \log(p(k|x,\theta))q(k|x). \quad (1)$$

The one-hot encoding is popular in classification tasks, which takes the form $q(y|x) = 1$ and $q(k|x) = 0$ for all $k \neq y$, for a sample x having class label y.

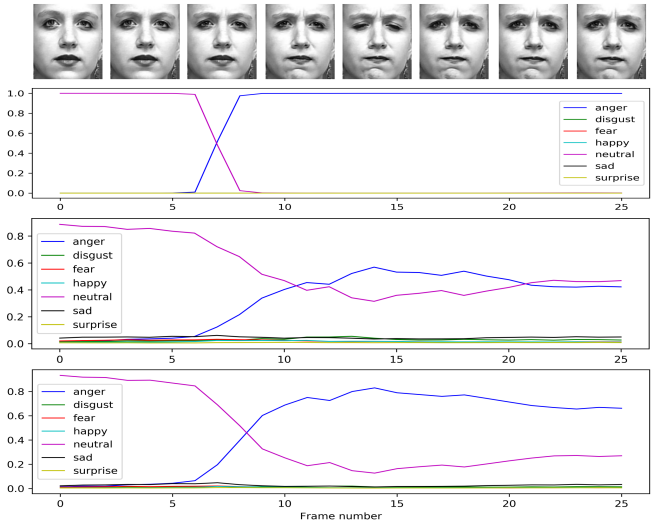


Fig. 1. Illustration of the effect of label smoothing and weak supervision. Top row: example expression-sequence in CK+; second row: prediction-scores without the use of label smoothing or unlabelled data; third row: prediction-scores with label smoothing *without* using unlabelled data; fourth row: prediction-scores when using label smoothing and weak supervision on unlabelled data. Best viewed in color.

Unlike object classification, expression categories are highly related, interconnected, as well as can occur simultaneously. For example, *happy* is different from *surprise*, however, both can occur jointly and simultaneously. In such cases, a model should provide prediction-scores of high probability for both expressions. However, one-hot vectors impose for the CNN to predict one of the class labels with high confidence, i.e., with probability 1. We noticed such FER-models, trained with one-hot encoding over-fit the data in most cases, i.e., they continuously generate high probability score for one of the expressions, irrespective of the presence of subtle or mixed emotions.

The limitations of over-fitted models, become evident in case of transition of expressions. Specifically, when facial expression changes from one expression to another, the probability score of an over-fitted model suddenly jumps from a negligible value (near to 0) to a large value (close to 1), or vice-versa in successive frames. Such an instance is demonstrated in Figure 1 (second row). However, we aim for our model to adapt to expression-intensity automatically. We used label smoothing and particularly a fraction of unlabelled data with replacement in each epoch to achieve this. The third and fourth row of Figure 1 demonstrate the effectiveness of

the above described proposed technique.

A. Label Smoothing

Label smoothing [23] seeks to replace one-hot vectors (i.e., 0 and 1 targets) with smoothed values (such as, 0.1 and 0.9, respectively), allowing for less confident predictions of the network, as well as to somewhat regularization of the model. To avoid large loss for erroneous samples, one-sided label smoothing is proposed in [20], where the positive labels are smoothed while setting the negative labels to 0. However, doing so will fail the model to adapt to mixed expressions. Therefore, we implemented the label smoothing as,

$$q'(k|x) = \begin{cases} 1-\epsilon, & k = y \\ \frac{\epsilon}{K-1}, & k \neq y, \end{cases} \tag{2}$$

where $\epsilon \in [0, 1]$ is the label smoothing hyperparameter. While setting $\epsilon = 0$ refers to one-hot encoding, setting ϵ a large value might result in learning a poor performing model. We note that label smoothing enables the model to be adaptable to unseen data.

B. Using Labelled and Unlabelled Data Simultaneously

A network trained with a limited-sized dataset might pitch into a local optimum. Assuming the network is already in its optimal state, the gradient descend algorithm would not change the network parameters, when unseen samples are used as training data along with their predicted probability scores as the ground-truth. For a relatively poor model, this helps the network to jump from the local optimum and reevaluate the state.

The proposed weakly supervised method uses a fraction of the unlabelled data along with the labelled samples to update the network. We used a self-training procedure inspired by [8], where the class labels of the unlabelled data are estimated using the network predictions. Unlike in [8], we use the predicted probability distribution as the ground-truth distribution.

Let us denote the labelled and unlabelled variables using $[.]_l$ and $[.]_u$ respectively (for example, X_l - labelled data, q_l - ground-truth distribution of labelled data, p_u - network prediction probabilities for unlabelled data, etc.). An initial model (θ^0) is trained with X_l until adequate performance is achieved. Further, the model parameters are updated in each epoch using a portion of both, X_l and X_u simultaneously. Maintaining the proper balance between the number of labelled and unlabelled data is very crucial for network performance. In our implementation, we randomly replace a fraction of X_l (typically $5 - 15\%$ of number of labelled samples in X_l) with the unlabelled data in each epoch. Moreover, incorrect predictions of unlabelled samples, that are used for training can deteriorate the model-performance after subsequent updates. In other words, since the model is not perfect in the beginning, there is a very high chance of having inaccurate predictions in the early training stages. Thus, the model might end up learning with inaccurate annotations in the early stages resulting in error accumulation as training proceeds. Therefore, we used the unlabelled data with high confidence scores ($\widehat{X_u}$), as suggested in [21]. The work flow of the proposed technique is shown in Figure 2.

After the t-th update, we obtained the prediction scores using θ^t for both X_l and X_u, denoted by p_l^t and p_u^t, respectively. The unlabelled data with high prediction scores are selected using a threshold value (τ), given by

$$\widehat{X_u^t} = \{x | x \in X_u \text{ and } \max_k p_u^t(k|x) > \tau\}; \quad \widehat{X_u^t} \subset X_u. \tag{3}$$

For the $(t+1)$-th iteration, the training set consists of X_l, with some of its data replaced by the samples from $\widehat{X_u^t}$. In other words, $X^{t+1} = \{X_l^{t+1}, X_u^{t+1}\}$, where $X_l^{t+1} \subset X_l$ and $X_u^{t+1} \subset \widehat{X_u^t}$ are selected randomly. Note that the ground-truth distribution of unlabelled data is additionally updated to its predicted probabilities, i.e., $q_u^{t+1}(k|x) = p_u^t(k|x)$. Thus, the model sees the unlabelled data with updated ground-truth distribution after each update. This allows the model to adapt to expression-intensities.

Choosing $\tau \approx 1$ refers to adding the unlabelled samples with absolute dominant class predictions. Such a model would not be able to adapt to moderate expression intensities. However, values of τ in the range $0.6 - 0.8$ is suitable, as it promises dominant class structure while adopting to moderate expression intensities.

1) Maintaining the Confidence of Labelled Samples: By performing successive label smoothing on labelled data, the model learns the expression intensities correctly. However, incorrect predictions can collapse the model-performance in subsequent iterations. It is important to maintain the prediction confidence of supervised data, while learning the necessary information from the unlabelled data. In order to achieve that, we scrutinize p_l^t after each iteration and force the model to rectify its prediction errors on supervised data (X_l). A simple way of doing so is to maintain a prediction score $> \alpha$ for the positive labels of supervised data, i.e., by using the following equations for $x \in X_l$ with its ground truth label y.

$$q_l^{t+1}(k|x) = \begin{cases} p_l^t(k|x), & \text{if } \{ \max_k p_l^t(k|x) > \alpha \\ & \text{and } \operatorname{argmax}_k p_l^t(k|x) = y \ \} \\ f(p_l^t(k|x)), & \text{otherwise} \end{cases} \tag{4}$$

$$\text{where} \quad f(p_l^t(k|x)) = \begin{cases} \alpha, & \text{if } k = y \\ \frac{1-\alpha}{K-1}, & \text{if } k \neq y. \end{cases} \tag{5}$$

The model is updated from θ^t to θ^{t+1} in a supervised manner using X^{t+1} and the corresponding updated ground-truth probabilities: q_l^{t+1} and q_u^{t+1}. This forces the model to generate closely similar probabilities every time it accepts a particular sample as input. Intuitively, the model will train itself to correctly classify the supervised data, while incorporating the variations from the unlabelled data into the model. Here the parameter α controls the prediction confidence of labelled data. Choosing the value of $\alpha \approx 1$ restricts the model to learn for definite dominant expressions.

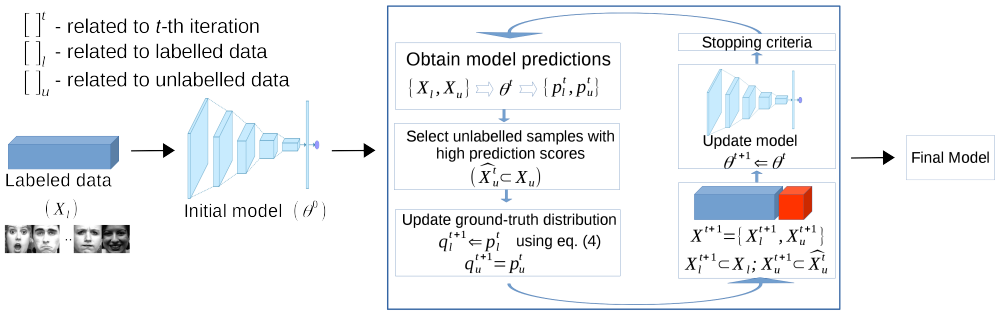


Fig. 2. Workflow of the proposed method.

However, as discussed before, data of different expression-intensities might have a similar label and the value of α should be in the range 0.6 to 0.9, in order to allow the model to fit such intensities with class dominance.

We use the decrease in average validation accuracy over last T iterations as the stopping criterion. In other words, we stop the training process if the trend of average validation accuracy starts increasing, i.e., $\overline{Acc_{val}^{t}} > \overline{Acc_{val}^{t-1}}$, where $\overline{Acc_{val}^{t}} = \frac{1}{T}\sum_{t=t-T+1}^{t} Acc_{val}^{t}$ is the average validation accuracy over previous T iterations. We used $T = 5$ in all experiments.

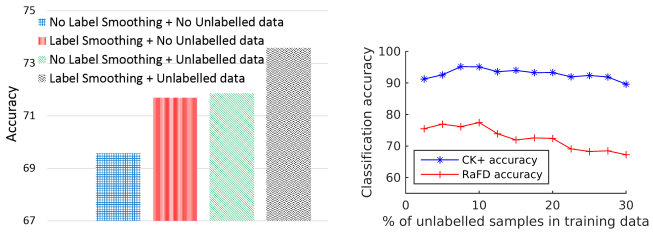
IV. Experiments and Results

A. Implementation Details

Datasets Experiments are conducted on four publicly available expression datasets, namely CK+ [24], RaFD [25], lifespan [26], and FER2013 [27]. In our experiments, we use 618, 1407, and 1027 samples from CK+, RaFD, and lifespan respectively for seven classes. Both RaFD and lifespan datasets contain static images, while CK+ contains image sequences. The image sequences in CK+ start from a neutral face and end with a peak of the respective expression. Therefore, we consider the first and last frame of each sequence as annotated with neutral and one of the six basic expressions respectively, while the intermediate frames of the corresponding sequence constitute the unlabelled data in our experiments. Lifespan database is particularly challenging, as it contains expressions from subjects of various age groups ranging from adolescents to elderly people. Moreover, it includes a range of expression-intensities from subtle to intense (as shown in Figure 3). FER2013 is an in-the-wild dataset containing train, validation, and test splits with 28709, 3589, and 3589 images in respective splits. In our experiments, the unlabelled data originates from the test-split during the experiment, unless specified.

Preprocessing steps involve face detection and face alignment, in order to position both eyes at a fixed distance parallel to the horizontal axis. The training set is augmented using slight zooming, horizontal flipping, less than 10% vertical and horizontal shifting, as well as rotating the images randomly in the range of ± 10 degrees.

Fig. 3. Samples from lifespan database illustrating the variation of expression intensity.



(a) (b)

Fig. 4. (a) Performance improvement in FER2013 with/without using label smoothing and unlabelled data. (b) Impact of varying the amount of unlabelled data on CK+ and RaFD.

Network We use the pre-trained VGG-Face model proposed by Parkhi *et al.* [11] initially introduced for face recognition. It consists of thirteen convolutional layers followed by two fully connected layers. We replace the last two fully connected layers by one fully connected layer with 128 neural units. Dropout is applied to the FC layer with a probability of 0.6. We set the softmax layer to the number of expression classes (in our case seven: anger, disgust, fear, happy, neutral, sadness, and surprise). We use adam optimizer [28] with the suggested weights $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 0.00001 in our model. All models are trained using a batch size of 32. After extensive experimentation, we conclude that fine-tuning the last convolutional and the fully connected layer is suitable for recognizing expressions.

Parameter Settings We conduct several experiments by varying ϵ from $0.02 - 0.25$, and varying both τ and α in the range $0.5 - 0.95$ to select the suitable values of corresponding parameters. Empirically we find $\epsilon = 0.1$, $\tau = 0.7$ and $\alpha = 0.7$ to be adequate for all the experiments, irrespective of the database type and mode of evaluation. Percentage of unlabelled data in the training set is another parameter which we discuss in section IV-C.

B. Effect of Label Smoothing and Unlabelled Data

Figure 4(a) illustrates the effect of label smoothing and unlabelled data. Here the baseline architecture is the VGG-Face model which achieves an accuracy of 69.57% on

FER2013 dataset. When label smoothing is applied to the baseline architecture, the model performance is improved by 2% demonstrating the effectiveness of label smoothing. This suggests that the use of repetitive supervised label smoothing improves the model performance by adapting to expression-intensities. Similar trends are observed by using only unlabelled data with no label smoothing. The performance gain achieved by applying each method independently is close. On the contrary, the accuracy is increased by 4% over the baseline performance when both the methods are combined. This shows that the use successive label smoothing and unlabelled data compliments each other and helps the model to learn the expression pattern in a better manner.

C. Selecting the Quantity of Unlabelled Data for the Training-set

Given the equal treatment of labelled and unlabelled sets during the model update, the choice of the amount of unlabelled data plays a crucial role in network training. While a too small number might not result in improvement in performance, a large amount might degrade the model performance. We employ 80%-20% of CK+ as train-test split and also perform cross-database evaluation on RaFD. Figure 4(b) illustrates that the CK+ test-set performance slightly degrades when using a larger amount of unlabelled data. However, the RaFD accuracy decreases continuously by increasing the percentage of unlabelled data. Therefore, we replace 10% of training data with unlabelled samples in all our experiments.

D. Cross-Database Evaluation

Figure 5 shows results related to the cross-database protocol (CK+ $\rightarrow$ RaFD, RaFD $\rightarrow$ CK+). For example, we train the model using the train-split of CK+, and test its performance on RaFD and the test-split of CK+ (see Figure 5(a)). We also report the model performance by, (i) using no unlabelled data, (ii) using unlabelled data from the test-split of the same dataset, and (iii) using unlabelled data from other dataset. We performed the experiments ten times and the average performance is reported. 80%-20% train-test split is used to obtain the results in Figure 5. As can be seen, the average test-split performance remains almost the same irrespective of the use of unlabelled samples. The unlabelled samples randomly replace the labelled data at each update. Thus, at every update, the network sees a particular labelled sample with a probability of 0.9 when using 10% unlabelled data. Since the network sees the labelled data repetitively, its performance is not affected substantially by the unlabelled data.

When trained on CK+ (see Figure 5(a)) with unlabelled data, the model-performance improves by 11% in RaFD. We observe that the use of unlabelled data from either CK+ or RaFD results in similar performances. Utilizing unlabelled images from CK+, the network sees varying expression-intensities and adapts to it. On the other hand, using unlabelled RaFD samples gradually makes the model aware of the test data, thus resulting in good accuracy. Similar conclusions can be drawn from Figure 5(b), where the model is trained on RaFD. The performance is improved by 7.5% using unlabelled images from CK+ sequences. We notice that in best case scenarios, the performance of the proposed model on CK+ has reached more than 90%.

Table I and Table II report classification results with respect to varying number of training samples. Significant classification accuracy has been obtained with merely 25% of the training data. Use of a larger labelled training set strikingly boosts the cross-database performance.

We observe that the cross-database performance on lifespan is lower, when trained with RaFD in comparison to CK+. This might be due to the presence of expressions of varying intensity in lifespan. However, when unlabelled samples from

TABLE I

CLASSIFICATION RESULTS USING CK+ DATABASE FOR TRAINING.

Test databases	Percentage of training data		
	25%	50%	80%
CK+ (test-set)	88.79%	91.29%	95.16%
RaFD	64.25%	65.25%	78.46%
lifespan	35.13%	40.51%	60.83%

TABLE II

CLASSIFICATION RESULTS USING RAFD DATABASE FOR TRAINING.

Test databases	Percentage of training data		
	25%	50%	80%
RaFD (test-set)	94.71%	97.24%	98.5%
CK+	79.8%	82.41%	86.64%
lifespan	28.24%	29.11%	34.96%

TABLE III

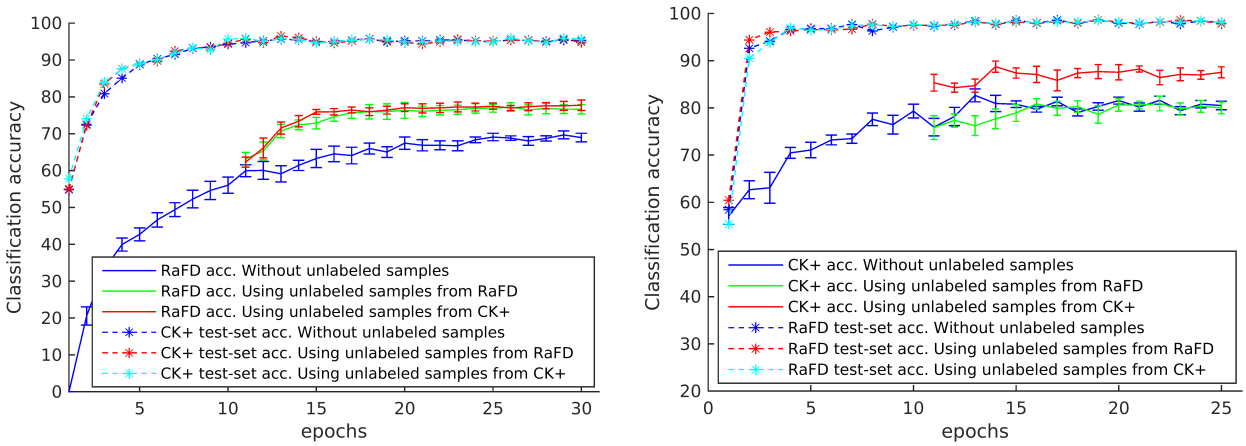
COMPARISON OF AVERAGE CLASSIFICATION ACCURACY ON CK+ DATABASE.

Methods	Validation settings	Accuracy
STM-ExpLet [29]	7 class	94.19
LOMo [30]	7 class	95.1
IACNN [6]	7 class	95.37
BDBN[3]	6 class	96.7
Facenet2expnet [5]	8 class	96.8
DTAGN [2]	7 class	97.25
PPDN [4]	6 class	97.3
facenet2expnet [5]	6 class	98.6
PPDN [4]	7 class	99.3
Proposed	7 class	**99.35**

TABLE IV

COMPARISON OF SEVEN CLASS CLASSIFICATION ACCURACY ON RAFD DATABASE.

Methods	Validation settings	Accuracy
Metric learning[31]	10 fold	95.95
W-CR-AFM [32]	train-test split	96.27
BAE-BNN-3[33]	5 fold	96.93
TLCNN+FOS[34]	4 fold	97.75
Carcagni *et al.*[35]	10 fold	98.5
Proposed	5 fold	**98.5**
Proposed	10 fold	**98.58**



(a) Trained on CK+ (b) Trained on RaFD

Fig. 5. Cross-database experiments show significant performance improvement. Best viewed in color.

TABLE V
COMPARISON OF AVERAGE ACCURACY ON FER2013 DATABASE.

Methods	Accuracy
Hand-crafted feature guided CNN [36]	61.86
AlexNet [37]	64.8
DNNRL [37]	70.6
ResNet [38]	72.4
VGG [38]	72.7
Ensemble of deep networks [39]	73.31
Alignment mapping networks + ensemble [39]	73.73
Single CNN [40]	71.47
Ensemble CNN [40]	73.73
Proposed	**73.58**

CK+ are used, the model learns the expression representation more accurately. Figure 1 depicts the smooth prediction scores, as well as the related image sequences, indicating how the model learned expression-intensities. Similar observations can be drawn on Fig. 5(b), where accuracy on CK+ is not affected when unlabelled data from RaFD is used. However, when the moderate intensity images (not coming from test sequences) from CK+ are used for training, the accuracy on CK+ improved by 10%. These observations demonstrate the adaptability of the network to the concerned task in a more regularized manner, instead of over-fitting for the database samples.

E. Comparison with Other Methods

The performance of the proposed approach is compared with other recent CNN-based methods and state-of-the-art results. However, we note that the validation strategy varies depending on the source-literature. Therefore, we report both, performance and validation settings in Table III and IV. We note that most of the literature used the last three images of the sequences of CK+ to report classification accuracy. For a fair comparison, we follow a similar protocol and considered 1236 labelled samples for CK+. Similar to our previous experiments, the unlabelled data consist of the intermediate frames excluding the labelled samples. The inclusion of more labelled samples in the training set improves the performance from 95.56% to 99.35% in CK+. As can be observed in Table III, this performance is higher than the previously reported results. Further, we notice that the inclusion of *neutral*-class decreased the performance of Facenet2expnet [5]. However, the proposed method achieves significant improvements, while using the neutral expression as one of the classes.

Similar observations are inferred on RaFD evaluation. Here the validation setting differs in literature in terms of the number of cross-validation folds. We perform both, 5-fold and 10-fold cross validation as shown in Table IV. Our approach outperform the previously reported performances in both cross-validation settings. Table V reports the performance of the proposed model on FER2013 dataset. As can be observed, our model achiev close to the state-of-the-art performance using a single framework.

V. CONCLUSIONS

In this paper, we propose a weakly supervised learning method that allows a CNN-model to adapt to different expression-intensities in addition to classifying them into discrete categories. Crucial in our approach is the utilized label smoothing and bootstrapping of a fraction of unlabelled samples, replacing labelled data for model-update, while maintaining the confidence level of the supervised data. Experiments conducted on four public datasets indicate (a) large performance-gain in cross-database evaluation, and (b) the self-adjustment of the network to different expression intensities.

REFERENCES

[1] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[2] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2983–2991.

[3] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805–1812.

[4] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *European conference on computer vision*. Springer, 2016, pp. 425–442.

[5] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 118–126.

[6] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 558–565.

[7] G. Papandreou, L.-C. Chen, K. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a dcnn for semantic image segmentation," *arXiv preprint arXiv:1502.02734*, 2015.

[8] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on Challenges in Representation Learning, ICML*, vol. 3, 2013, p. 2.

[9] S. Hong, H. Noh, and B. Han, "Decoupled deep neural network for semi-supervised semantic segmentation," in *Advances in neural information processing systems*, 2015, pp. 1495–1503.

[10] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, 2017.

[11] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[12] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[13] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1717–1724.

[14] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*. ACM, 2015, pp. 443–449.

[15] A. Coates and A. Y. Ng, "Learning feature representations with k-means," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 561–580.

[16] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[17] J. Zhao, M. Mathieu, R. Goroshin, and Y. Lecun, "Stacked what-where autoencoders," *arXiv preprint arXiv:1506.02351*, 2015.

[18] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 3546–3554.

[19] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[20] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[21] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," 2005.

[22] S. Novotney, R. Schwartz, and J. Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4297–4300.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[24] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.

[25] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[26] M. Minear and D. C. Park, "A lifespan database of adult facial stimuli," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 4, pp. 630–633, 2004.

[27] I. J. Goodfellow, D. Erhan, P. L. Carrier *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1749–1756.

[30] K. Sikka, G. Sharma, and M. Bartlett, "Lomo: Latent ordinal model for facial analysis in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5580–5589.

[31] B. Jiang and K. Jia, "Robust facial expression recognition algorithm based on local metric learning," *Journal of Electronic Imaging*, vol. 25, no. 1, p. 013022, 2016.

[32] B.-F. Wu and C.-H. Lin, "Adaptive feature mapping for customizing deep learning based facial expression recognition model," *IEEE Access*, vol. 6, pp. 12 451–12 461, 2018.

[33] W. Sun, H. Zhao, and Z. Jin, "An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks," *Neurocomputing*, vol. 267, pp. 385–395, 2017.

[34] Y. Zhou and B. E. Shi, "Action unit selective feature maps in deep networks for facial expression recognition," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 2031–2038.

[35] P. Carcagnì, M. Coco, M. Leo, and C. Distante, "Facial expression recognition and histograms of oriented gradients: a comprehensive study," *SpringerPlus*, vol. 4, no. 1, p. 645, 2015.

[36] G. Zeng, J. Zhou, X. Jia, W. Xie, and L. Shen, "Hand-crafted feature guided deep learning for facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 423–430.

[37] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, and D. Tao, "Deep neural networks with relativity learning for facial expression recognition," in *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2016, pp. 1–6.

[38] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: state of the art," *arXiv preprint:1612.02903*, 2016.

[39] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, "Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 48–57.

[40] Y. Gan, J. Chen, and L. Xu, "Facial expression recognition boosted by soft label with a diverse ensemble," *Pattern Recognition Letters*, vol. 125, pp. 105–112, 2019.